Memo

Date:	Friday, December 24, 2021
Project:	Comanche Station Groundwater Monitoring – Statistical Analysis
To:	Jennifer McCarter (PSCo)
From:	May Raad, M.Sc, PStat®, Principal Statistician (HDR)
Subject:	Response to Comanche Compliance Issues

The following responses are provided by May Raad, Principal Statistician at HDR since 2006.

May Raad has accredited designations of PStat® from the American Statistical Association and P.Stat.from the Statistical Society of Canada. May has over 30 years of experience providing statistical and data analysis and reporting to natural resources agencies, power and water utilities, transportation agencies, and government agencies. She has the exciting role of leading HDR's statistics and data analysis practice in the engineering sector. Assessing and forecasting trends in water quality data is a highly specialized area in statistics. During the last decade, she has applied advanced statistical models and methodologies to analyze groundwater quality samples for power utilities and government agencies across North America.

Question 1:

§257.93(a) and §257.93(g)(1)- The statistical procedures must be "designed to ensure monitoring results that provide an accurate representation of groundwater quality at the background wells," including that "the statistical method used to evaluate groundwater monitoring data shall be appropriate for the distribution of constituents." More information is needed to determine if the distributional models used in calculation of the background threshold values (BTV) are appropriate. On page 1 in the "Attachment 1 to November 4, 2021 letter" under "Identify Best Fit" it states: "When data sets can be fit to multiple distribution models, the following hierarchy is applied for selecting the appropriate distribution: I. gamma, II. lognormal, III. normal" However, normality testing data were not submitted, nor was justification provided for this automated hierarchy of selecting gamma over normal (or vice-versa). Examples of objective data to justify selection of gamma over normal distribution model include comparative normality test scores and elevated skewness scores.

Please submit normality and skewness testing results for each data set……

Xcel is providing the normality and skewness testing results in the following excel files:

1. ProUCL Output - GOF Tests.xlsx – provides all the GOF test results output by ProUCL software for data with and without (nondetects) NDs, which tests for normality,

lognormality, and gamma, for each consistent of concern at the Comanche Bottom Ash Pond and Comanche Landfill.

2. Skewness.xlsx – consists of tables of the sample sizes, level of censorship, distribution, and skewness levels for each constituent of concern at the Comanche Bottom Ash Pond and Comanche Landfill.

The ProUCL Technical Guide (USEPA, 2015) defines skewness as a function of the sd of logged data. This measure of skewness is not only applied to lognormally distributed data sets but also other data sets with positive values. The skewness levels provided in the table below have been defined in Table 2-1 as well as Table A-7 of the ProUCL Technical Guide.

Table 1: Skewness as a Function of sd

SD of logged data	Skewness
SD < 0.5	Symmetric to mild skewness
0.5 <= SD < 1.0	Mild skewness to moderate skewness
1.0 <= SD < 1.5	Moderate skewness to high skewness
1.5 <= SD < 2.0	High skewness
2.0 <= SD < 3.0	Very high skewness (moderate probability of outliers and/or multiple populations)
SD >= 3.0	Extremely high skewness (high probability of outliers and/or multiple populations)

…..and provide justification for the use of an automated hierarchy for selecting one distributional model over another apart from objective data.

The hierarchy (i.e., I. gamma, II. lognormal, III. normal) applied for selecting the appropriate distribution to compute decision-making statistics (e.g., UPLs and UTLs) when data sets can be fit to multiple distribution models is based on the recommendations in the ProUCL Technical Guide.

ProUCL outputs its statistical methods for comparison, academic, and research purposes, in the anticipation that it will help decision makers make more informative and defensible decisions.

The use and applicability of a statistical method depend on:

- Data size
- Level of censorship
- Data skewness
- Data distribution

The ProUCL Technical Guide emphasizes that it is well known in the literature that environmental data are often right-skewed and skewed distributions, such as the lognormal and gamma, are routinely used to model such data. In particular, the gamma model is commonly used to model environmental data. Gibbons and Coleman (2001, pp. 34–47) noted that the use of a gamma distribution is more appropriate than a normal distribution when variability and concentration are related, as in the case of many environmental constituents. Millard (2013)

states some EPA guidance documents strongly recommend using a gamma distribution for environmental data rather than a lognormal model.[1]

Practitioners tend to use the central limit theorem (CLT) or Student's t-statistic (normal) based BTVs for samples of sizes 25-30 (large sample rule-of-thumb to use CLT). ProUCL Technical Guide states, "However, this rule-of-thumb does not apply to moderately skewed to highly skewed data sets, specifically when the sd of the log-transformed data starts exceeding 1". The large sample requirement for following an approximate normal distribution increases with skewness.[2]

[1] Examples of such references in ProUCL as to the appropriateness or advantages of the gamma distribution can be read in the Executive Summary on page vi "*The use of a parametric lognormal distribution on a lognormally distributed data set yields unstable impractically large UCLs values, especially when the standard deviation (sd) of the log-transformed data becomes greater than 1.0 and the data set is of small size less than 30-50. Many environmental data sets can be modeled by a gamma as well as a lognormal distribution. The use of a gamma distribution on gamma distributed data sets tends to yield UCL values of practical merit. Therefore, the use of gamma distribution based decision statistics such as UCLs, UPLs, and UTLs cannot be dismissed by stating that it is easier (than a gamma model) to use a lognormal model to compute these upper limits*.",

Page 38: *"It is further stated in Helsel (2012a) that ProUCL prefers the gamma distribution because it downplays outliers as compared to the lognormal. This argument can be turned around - in other words, one can say that the lognormal is preferred by practitioners who want to inflate the effect of the outlier. Setting this argument aside, we prefer the gamma distribution as it does not transform the variable so the results are in the same scale as the collected data set. As mentioned earlier, log-transformation does appear to be simpler but problems arise when practitioners are not aware of the pitfalls (e.g., Singh and Ananda 2002; Singh, Singh, and Iaci 2002) associated with the use of lognormal distribution*.", and;

Page 138, "*Furthermore, when using a gamma distribution to compute decision statistics such as a UCL of the mean, one does not have to transform the data and back-transform the resulting UCL into the original scale*."

[2] *It is noted that even for skewed data sets, practitioners tend to use the CLT or Student's t-statistic based UCLs of mean for "large" sample sizes of 25-30 (rule-of-thumb to use CLT). However, this rule-of-thumb does not apply for moderately to highly skewed data sets, specifically when σ (standard deviation of the log-transformed data) starts exceeding 1. The large sample size requirement associated with the use of the CLT depends upon the skewness of the data distribution under consideration. The large sample requirement associated with CLT for the sample mean to follow an approximate normal distribution increases with the data skewness; and for highly skewed data sets, even samples of size greater than (>)100 may not be large enough for the sample mean to follow an approximate normal distribution. For moderately skewed to highly skewed environmental data sets, as expected, UCLs based on the CLT and the Student's t-statistic fail to provide the desired coverage of the population mean even when the sample sizes are as large as 100 or more. These facts have been verified in the published simulation experiments conducted on positively skewed data sets (e.g., Singh, Singh, and Engelhardt, 1997; Singh, Singh, and Iaci, 2002); some graphs showing the simulation results are provided in Appendix B.)."* ProUCL, p. 1.

Tables in Attachments 1a and 1b show that the gamma distributional model was selected for constituents when both the normal and gamma distributional models were only described concurrently as "best fit." The resulting gamma-based BTVs selected for use were often significantly higher than the normal-based BTVs (e.g. gamma-Fluoride = 735 mg/L vs. normal-Fluoride = 236 mg/L).

It can be observed (in the tables provided) that when the skewness levels are 'symmetric to mild skewness' the gamma-based BTVs are very similar to the normal-based BTVs. As the skewness levels increase, then the gamma-based BTVs appear to be greater than the normal-based BTVs.

There was only one case where both the normal and gamma distributional models were only described concurrently as "best fit" and this was for fluoride at the Bottom Ash Pond.

We did go back to the eight collected background samples at W-2A for fluoride for the Bottom Ash Pond because of EPA's question regarding the fluoride UTL. The skewness level as defined by ProUCL is high at 1.73. As the skewness levels increase, then the gamma-based BTVs appear to be greater than the normal-based BTVs. With respect to the fluoride example, the differences between the Wilson-Hilferty (WH) and Hawkins-Wixley (HW) UTLs are larger than one would expect given that the transformed values follow an approximate normal distribution.

From our experience, notable differences happen when there is an outlier in the dataset, or the dataset is highly skewed. No statistical outliers were observed in the preliminary data analysis for fluoride, which uses Dixon's outlier test to identify outliers in the detected data. Upon further investigation of the data, we found that the laboratory reported value of 170 milligrams per litre (mg/l) sampled on December 2, 2020 was assigned a method detection limit (MDL) of 170 mg/l due to dilutions; however, every other date the MDL was much lower. This suggests that the lab reported fluoride from a 1000X dilution, which rendered the MDL high on that date. This concentration was the maximum value for fluoride in the background dataset. After further review and discussions with the laboratory, we recommend excluding the value of 170 mg/l from the dataset as the dilution was too high to yield a reliable result. Using our methodology for sample size less than eight, we recommend nonparametric methods, hence the maximum detected value is used to represent the UPL and UTL for fluoride, i.e., 99 mg/l (USEPA. 1992).

These large differences become even more concerning considering the higher level of uncertainty surrounding the performance of the gamma-based BTVs used in ProUCL versus the highly-studied and well-documented normal upper prediction limit (UPL) and upper tolerance limit (UTL) (e.g. p.108 in ProUCL Version 5.1.002 Technical Guide - "Note: It should be pointed out that the performance of gamma UTLs and gamma UPLs based upon these HW and WH approximations is not well-studied and documented").

Even though methods exist to compute 95% (upper confidence limits) UCLs of the mean, upper prediction limits (UPLs) and upper tolerance limits (UTLs) based upon gamma distributed data sets, those methods have not become popular due to the computational complexity and/or the lack of availability in commercial software packages. Despite the better performance (in terms of coverage and stability) of the decision-making statistics based upon a gamma distribution, some

practitioners tend to dismiss the use of gamma distribution-based decision statistics by not acknowledging them and/or stating that the use of a lognormal distribution is easier to compute the various upper limits. However, one should not compromise the accuracy and defensibility of estimates and decision statistics by using easier methods. Computation of defensible estimates and decision statistics taking the sample size and data skewness into consideration is always recommended (USEPA 2015).

The ProUCL Technical Guide references studies that have demonstrated that UCLs based upon the CLT and Student's t statistics fail to provide the desired 95% coverage of the population mean for small sample sizes, even ones as large as 100, for skewed distributions. Moreover, the properties of the CLT and Student’s t-statistic are unknown when NDs with varying DLs are present in a dataset. The use of parametric lognormal distribution on a lognormally distributed data set tend to yield unstable impractically large UCL values, especially when the sd of the log-transformed data is greater than 1 and the data set is of small size (less than 30 to 50). Similar patterns are expected in the behavior and properties of the various other upper limits (e.g., UTLs, UPLs) used in the decision-making processes of the USEPA.

Generally, the use of a gamma distribution on gamma distributed data sets yields reliable and stable UCL values of practical merit. Therefore, the use of gamma distribution-based decision statistics such as UCLs, UPLs, and UTLs should not be dismissed just because it is easier to use a lognormal or normal model. The gamma distribution is suggested for computing BTV estimates since it accounts for data skewness. The advantages of computing the gamma distribution-based decision statistics are discussed throughout the ProUCL Technical Guide.

ProUCL has incorporated BTVs based upon normal approximation to the gamma distribution, which are based upon Wilson-Hilferty (WH) and Hawkins-Wixley (HW) approximations. Other references to this methodology can be found in recent literature. Krishnamoorthy, Mathew, and Mukherjee (2008) present simple methods for finding two-parameter gamma tolerance limits and prediction limits, which are essentially normal-based methods applied to cube-root-transformed samples. Krishnamoorthy and Wang (2016) describe how to find confidence limits for the mean and an upper percentile, and upper prediction limits for the mean of a future sample from a gamma distribution for censored and uncensored cases. Monte Carlo simulation studies indicate that the methods are accurate for estimating the mean and percentile and for predicting the mean of a future sample as long as the percentage of NDs is not too large.

Question 2

§257.93(g)(4)- The selection of a prediction limit method or tolerance limit method in detection or assessment monitoring "..shall be such that this approach is at least as effective as any other approach in this section for evaluating groundwater data." The practice of "averaging of the results over the multiple methods" to create a "pooled estimate" of the background parameter (as described on page 1 of the "Attachment 1 to November 4, 2021 letter" and Attachments 1a and 1b) does not meet the requirement of this rule.

For example, Attachment 1a shows two distinct UTLs for fluoride (WH-UTL = 614 mg/L; HW-UTL = 857 mg/L), with the final BTV for fluoride as an arithmetic average of the two (i.e. 735 mg/L). The final, averaged UTL of 735 mg/L is 121 mg/L higher than the WH-UTL of 614 mg/L. Therefore, if the final, averaged UTL of 735 mg/L is used as a background-based groundwater protection standard (GPS) in assessment it would have lower statistical power to detect upward changes in fluoride as compared to use of the WH-UTL of 614 mg/L, and thus the averaged UTL would not be "as effective" as the WH-UTL.

Additionally, on top of the uncertainty regarding performance of gamma UTLs described in 1) above, the performance of "pooled" or "averaged" gamma UTLs is even more unstudied, uncertain and thus inappropriate.

HDR has provided a pooled methodology in situations where multiple models are provided by ProUCL. The averaging of competing models used to estimate the same statistic corrects for bias inherent in each model. Our approach is at least as effective as the standards required by the CCR Rule and §257.93(g)(4).

The CCR Final Rule documented that a sample of at least eight is sufficient (see discussion on page 21401; Volume 80 of the Federal Register). A sample size of eight does provide data from which a sample variance and statistical results can be produced; however, it does not technically provide for an unbiased representation of the true underlying distribution of constituent concentrations at a site. The averaging of models or expert opinions has been shown to produce superior results in disciplines such as economics (Ouchi, F., 2004), statistics (Lavancier and Rochet, 2013) and data mining and machine learning (Nisbet, Elder, Miner, 2009; Giovanni, Elder, 2010). HDR recommends this approach to reduce the chance of any one method, incorrectly chosen due to too small a sample, producing highly biased estimates.

Please note that the intention of averaging is not to bias an estimate such as a UPL or UTL in any one direction. It is direction 'indifferent' and over all constituents and samples using this approach, the chances of average higher error rates in the statistics are lowered.

The Wilson-Hilferty (WH) and Hawkins-Wixley (HW) methods to estimate UPLs or UTLs for gamma distributions have been well-studied. As recently as 2021, these methods are being used to update the guidance manual "Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory Approved Guideline – Third Edition, EP28-A3C".[3]

[3] HDR contacted Dr. Douglas Hawkins (key author of the HW method) on December 16, 2021 to inquire the usage of his methodology described in his and R. A. Wixley's 1986 paper. He indicated the working group (in which he is a member) organized to update the CLSI guidelines from version EP-28 to EP44 are incorporating both WH and HW power transformations to aid in developing limits for reference intervals.

The basis of the two methods is a power transformation to transform the data to follow a normal distribution. Once normalized, the formula for UPL or UTL uses the same formula as what one would use under the well-studied normal distribution assumptions to estimate a UPL or UTL, that is, the sample average plus the K tolerance factor multiplied by the standard error of the sample. The output from the calculation is then transformed back to the original scale using the inverse of the power transformation.

The WH method uses the cube root (1/3) power transformation and the HW use the fourth root transformation (1/4). Both transformations perform well to model a normal distribution (Krishnamoorthy, Mathew & Mukherjee 2008). To choose one over the other based on small sample sizes taken over 1 to 2 years may be a risk and does not necessarily provide certainty one expects. The averaging of the competing methods offered in ProUCL is done to mitigate that uncertainty.

Regarding the question about which UTL produced by the WH and HW methods has more power, we note that the UTLs for assessment monitoring are used as proxy GPS if no MCLs are published or if the GPS are higher than their respective MCLs. As recommended by the Unified Guidance (USEPA 2009), the UTLs should be treated as fixed numbers and not statistical tests. This means that statistical power does not factor into the selection of the UTL. The Unified Guidance (UG) states the following:

> "Existing background levels may also exceed a fixed GPS.[4] In these cases, a background standard can be constructed using an upper tolerance interval on background with 95% confidence and 95% coverage. The standard will then represent a reasonable upper bound on background and an achievable target for compliance and remediation testing."

(We did provide our methodology to address sufficient power to detect an SSI using UPLs in our previous response based on a site-wide false positive rate of 10% and verification samples. (November 5, 2021).)

Historically, practitioners would simply use multiples of the background average to denote a GPS (e.g., two times the sample average). The UG addresses this approach in this manner:

> "However, this approach may not fully account for natural variation in background levels and lead to higher than expected false positive rates. If the GPS were to be set at the historical background sample mean, even higher false positive rates would occur during compliance monitoring, and demonstrating corrective action compliance becomes almost impossible."

The UG continues on the merits of using the background sample and variability to identify a realistic proxy for the GPS.

> "…, an upper tolerance limit based on both background sample size and sample variability is recommended for identifying the background GPS at a suitably high enough level above current background to allow for reversal of the test hypotheses. Although a

[4] Note that the UG uses GWPS in the exact quotes. For consistency, Xcel used GPS in place of GWPS to match language in Question 1.

somewhat arbitrary choice, a GPS based on this method allows for a variety of confidence interval tests".

In assessment monitoring and corrective action monitoring, the statistical testing by means of confidence intervals tests comes into play when downgradient samples are used to produce lower confidence limits (LCL) (assessment monitoring) or upper confidence limits (UCL) (corrective action compliance). These tests are one sample tests, or in other words, testing if a sample statistic is less than a fixed number (i.e., the GPS) while accounting for the variability in the sample.

Statistical power is increased in these tests with every round of sampling, as the monitoring dataset increases in size.

As the UG has indicated, the selection of the background GPS may seem arbitrary but using an appropriate distribution that best describes natural variability of concentrations and provides upper limit estimates will yield a practical GPS that has the chance of correctly flagging SSLs in assessment monitoring and achieving compliance in corrective action.

Our averaging of the two potential UTLs (WH UTL and HW UTL) is done precisely to provide a practical GPS that describes upper limits of variability bounded by these methods.

References:

Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory Approved Guideline – Third Edition, EP28-A3C, https://clsi.org/standards/products/method-evaluation/documents/ep28/, accessed Dec. 15, 2021.

Millard, S. (2013). EnvStats: An R Package for Environmental Statistics. Springer, New York. ISBN 978-1-4614-8455-4

Gibbons, R. D., & Coleman, D. D. (2001). Statistical methods for detection and quantification of environmental contamination. Hoboken, NJ: Wiley.

Giovanni, S., & Elder, J. (2010). "Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions.", Synthesis Lectures on Data Mining and Knowledge Discovery, edited by Robert Grossman, Morgan and Claypool Publishers, 2010

Hawkins, D. M., and Wixley, R. A. J. 1986. A Note on the Transformation of Chi-Squared Variables to Normality. The American Statistician, 40, 296–298.

K Krishnamoorthy, Thomas Mathew & Shubhabrata Mukherjee (2008) Normal-Based Methods for a Gamma Distribution, Technometrics, 50:1, 69-78, DOI: 10.1198/004017007000000353

Krishnamoorthy, K. & Wang, Xiao. (2016). Fiducial confidence limits and prediction limits for a gamma distribution: Censored and uncensored cases. Environmetrics. 27. 10.1002/env.2408.

Lavancier, F., P. Rochet. (2013). A general procedure to combine estimators. Computational Statistics & Data Analysis 94, December, 2013.

Nisbet, R., Elder, J. & Miner, G. (2009). Handbook of Statistical Analysis and Data Mining, Academic Press, Burlington, MA

Ouchi, F. (2004). Literature Review on the Use of Expert Opinion in Probabilistic Risk Analysis. World Bank Policy Research Working Paper 3201

Singh, A., A.K. Singh, and R.J. Iaci. (2002). Estimation of the Exposure Point Concentration Term Using a Gamma Distribution. EPA/600/R-02/084. October 2002. Technology Support Center for Monitoring and Site Characterization, Office of Research and Development, Office of Solid Waste and Emergency Response, U.S. Environmental Protection Agency, Washington, D.C.

USEPA. 1992. Supplemental Guidance to RAGS: Calculating the Concentration Term. Publication EPA 9285.7-081, May 1992.

U.S. Environmental Protection Agency (EPA). 2006a, Guidance on Systematic Planning Using the Data Quality Objective Process, EPA QA/G-4, EPA/240/B-06/001. Office of Environmental Information, Washington, DC. Download from: http://www.epa.gov/quality/qs-docs/g4-final.pdf

U.S. Environmental Protection Agency (EPA). 2006b. Data Quality Assessment: Statistical Methods for Practitioners, EPA QA/G-9S. EPA/240/B-06/003. Office of Environmental Information, Washington, DC. Download from: http://www.epa.gov/quality/qs-docs/g9s-final.pdf

USEPA, 2009. Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities: Unified Guidance. Office of Resource Conservation and Recovery, Program Implementation and Information Division, USEPA, EPA 530/R-09-007, 2009.

USEPA, October 2015. ProUCL Version 5.1 Technical Guide. USPEA, Office of Research and Development, Washington D.C. EPA/600/R-07/041.

USEPA, 40 CFR Parts 257 and 261, (2015) Hazardous and Solid Waste Management System; Disposal of Coal Combustion Residuals From Electric Utilities; Final Rule

Wilson, E. B., and Hilferty, M. M. (1931), “The Distribution of Chi-Squares,” Proceedings of the National Academy of Sciences, 17, 684–688.